# 12 Best Large Language Models (LLMs) in 2024

beebom.com/best-large-language-models-llms/

If you are discussing technology in 2024, you simply can't ignore trending topics like Generative AI and large language models (LLMs) that power AI chatbots. After the release of ChatGPT by OpenAI, the race to build the best LLM has grown multi-fold. Large corporations, small startups, and the open-source community are working to develop the most advanced large language models. So far, more than hundreds of LLMs have been released, but which are the most capable ones? To find out, follow our list of the best large language models (proprietary and open-source) in 2024.

## 1. GPT-4

The GPT-4 model by OpenAI is the best AI large language model (LLM) available in 2024. Released in March 2023, the GPT-4 model has showcased tremendous capabilities with complex reasoning understanding, advanced coding capability, proficiency in multiple academic exams, skills that exhibit human-level performance, and much more

In fact, it's the first multimodal model that can **accept both texts and images** as input. Although themultimodal ability has not been added to ChatGPT yet, some users have got access via Bing Chat, which is powered by the GPT-4 model.

> Big new AI thing: Microsoft Bing (which uses GPT-4 in Creative Mode), accepts images as input.
>
> The results are impressive. I fed it a meme, it could understand context & read text! A new dimension of AI use just opened up. So expect a flood of AI Twitter influencer threads… pic.twitter.com/pshP6J44tK— Ethan Mollick (@emollick) June 21, 2023

Apart from that, GPT-4 is one of the very few LLMs that has addressed hallucination and improved factuality by a mile. In comparison to ChatGPT-3.5, the GPT-4 model scores **close to 80% in factual evaluations** across several categories. OpenAI has also worked at great lengths to make the GPT-4 model more aligned with human values using Reinforcement Learning from Human Feedback (RLHF) and adversarial testing via domain experts.

The GPT-4 model has been trained on a massive 1+ trillion parameters and supports a maximum context length of **32,768 tokens**. Until now, we didn't have much information about GPT-4's internal architecture, but recently George Hotz of The Tiny Corp revealed GPT-4 is a **mixture model** with 8 disparate models having 220 billion parameters each. Basically, it's not one big dense model, as understood earlier.

Finally, you can use ChatGPT plugins and browse the web with Bing using the GPT-4 model. The only few cons are that it's slow to respond and the inference time is much higher, which forces developers to use the older GPT-3.5 model. Overall, the OpenAI GPT-4 model is by far the best LLM you can use in 2024, and I strongly recommend subscribing to ChatGPT Plus if you intend to use it for serious work. It costs $20, but if you don't want to pay, you can use ChatGPT 4 for free from third-party portals.

***Check out GPT-4***

## 2. GPT-3.5

After GPT 4, OpenAI takes the second spot again with GPT-3.5. It's a general-purpose LLM similar to GPT-4 but lacks expertise in specific domains. Talking about the pros first, it's an **incredibly fast model** and generates a complete response within seconds.

Whether you throw creative tasks like writing an essay with ChatGPT or coming up with a business plan to make money using ChatGPT, the GPT-3.5 model does a splendid job. Moreover, the company recently released a larger 16K context length for the GPT-3.5-turbo model. Not to forget, it's also free to use and there are no hourly or daily restrictions.

That said, its biggest con is that GPT-3.5 **hallucinates a lot** and spews false information frequently. So for serious research work, I won't suggest using it. Nevertheless, for basic coding questions, translation, understanding science concepts, and creative tasks, the GPT-3.5 is a good enough model.

In the HumanEval benchmark, the GPT-3.5 model scored 48.1% whereas GPT-4 scored 67%, which is the highest for any general-purpose large language model. Keep in mind, GPT-3.5 has been trained on 175 billion parameters whereas GPT-4 is trained on more than 1 trillion parameters.

## 3. PaLM 2 (Bison-001)

Next, we have the PaLM 2 AI model from Google, which is ranked among the best large language models of 2024. Google has focused on commonsense reasoning, formal logic, mathematics, and advanced coding in 20+ languages on the PaLM 2 model. It's being said that the largest PaLM 2 model has been **trained on 540 billion parameters** and has a maximum context length of 4096 tokens.

Google has announced four models based on PaLM 2 in different sizes (Gecko, Otter, Bison, and Unicorn). Of which, Bison is available currently, and it scored 6.40 in the MT-Bench test whereas GPT-4 scored a whopping 8.99 points.

Google Bard running on PaLM 2

That said, in reasoning evaluations like WinoGrande, StrategyQA, XCOPA, and other tests, PaLM 2 does a remarkable job and outperforms GPT-4. It's also a **multilingual model** and can understand idioms, riddles, and nuanced texts from different languages. This is something that other LLMs struggle with.

One more advantage of PaLM 2 is that it's very quick to respond and offers three responses at once. You can follow our article and test the PaLM 2 (Bison-001) model on Google's Vertex AI platform. As for consumers, you can use Google Bard which is running on PaLM 2.

***Check out PaLM 2***

## 4. Claude v1

In case you are unaware, Claude is a powerful LLM developed by Anthropic, which has been backed by Google. It has been co-founded by former OpenAI employees and its approach is to build AI assistants which are **helpful, honest, and harmless**. In multiple benchmark tests, Anthropic's Claude v1 and Claude Instant models have shown great promise. In fact, Claude v1 performs better than PaLM 2 in MMLU and MT-Bench tests.

Claude via Slack

It's close to GPT-4 and scores 7.94 in the MT-Bench test whereas GPT-4 scores 8.99. In the MMLU benchmark as well, Claude v1 secures 75.6 points, and GPT-4 scores 86.4. Anthropic also became the first company to offer **100k tokens as the largest context window** in its Claude-instant-100k model. You can basically load close to 75,000 words in a single window. That's absolutely crazy, right? If you are interested, you can check out our tutorial on how to use Anthropic Claude right now.

***Check Out Claude v1***

## 5. Cohere

Cohere is an AI startup founded by former Google employees who worked on the Google Brain team. One of its co-founders, Aidan Gomez was part of the "*Attention is all you Need*" paper that introduced the Transformer architecture. Unlike other AI companies, Cohere is here for enterprises and solving generative AI use cases for corporations. Cohere has a number of models from small to large — having **just 6B parameters** to large models trained on 52B parameters.

The recent Cohere Command model is **winning praise for its accuracy and robustness**. According to Standford HELM, the Cohere Command model has the highest score for accuracy among its peers. Apart from that, companies like Spotify, Jasper, HyperWrite, etc. are all using Cohere's model to deliver an AI experience.

In terms of pricing, Cohere charges **$15 to generate 1 million tokens** whereas OpenAI's turbo model charges $4 for the same amount of tokens. Nevertheless, in terms of accuracy, it's better than other LLMs. So if you run a business and looking for the best LLM to incorporate into your product, you can take a look at Cohere's models.

***Check Out Cohere***

## 6. Falcon

Falcon is the first open-source large language model on this list, and it has outranked all the open-source models released so far, including LLaMA, StableLM, MPT, and more. It has been developed by the Technology Innovation Institute (TII), UAE. The best thing about Falcon is that it has been **open-sourced with Apache 2.0 license**, which means you can use the model for commercial purposes. There are no royalties or restrictions either.

So far, the TII has released two Falcon models, which are **trained on 40B and 7B parameters**. The developer suggests that these are raw models, but if you want to use them for chatting, you should go for the Falcon-40B-Instruct model, fine-tuned for most use cases.

The Falcon model has been primarily trained in English, German, Spanish, and French, but it can also work in Italian, Portuguese, Polish, Dutch, Romanian, Czech, and Swedish languages. So if you are interested in open-source AI models, first take a look at Falcon.

***Check out Falcon***

## 7. LLaMA

Ever since LLaMA models leaked online, Meta has gone all-in on open-source. It officially released LLaMA models in various sizes, from **7 billion parameters to 65 billion parameters**. According to Meta, its LLaMA-13B model outperforms the GPT-3 model from OpenAI which has been trained on 175 billion parameters. Many developers are using LLaMA to fine-tune and create some of the best open-source models out there. Having said that, do keep in mind, LLaMA has been released for research only and can't be used commercially unlike the Falcon model by the TII.

Talking about the LLaMA 65B model, it has shown amazing capability in most use cases. It ranks among the top 10 models in Open LLM Leaderboard on Hugging Face. Meta says that it has not used any proprietary material to train the model. Instead, the company has used **publicly available data** from CommonCrawl, C4, GitHub, ArXiv, Wikipedia, StackExchange, and more.

Simply put, after the release of the LLaMA model by Meta, the open-source community saw rapid innovation and came up with novel techniques to make smaller and more efficient models.

***Check out LLaMA***

## 8. Guanaco-65B

Among the several LLaMA-derived models, Guanaco-65B has turned out to be the best open-source LLM, just after the Falcon model. In the MMLU test, it scored 52.7 whereas the Falcon model scored 54.1. Similarly, in the TruthfulQA evaluation, Guanaco came up with a 51.3 score and Falcon was a notch higher at 52.5. There are four flavors of Guanaco: 7B, 13B, 33B, and 65B models. All of the models have been **fine-tuned on the OASST1 dataset** by Tim Dettmers and other researchers.

As to how Guanaco was fine-tuned, researchers came up with a **new technique called QLoRA** that efficiently reduces memory usage while preserving full 16-bit task performance. On the Vicuna benchmark, the Guanaco-65B model outperforms even ChatGPT (GPT-3.5 model) with a much smaller parameter size.

The best part is that the 65B model has trained on a single GPU having 48GB of VRAM in just 24 hours. That shows how far open-source models have come in reducing cost and maintaining quality. To sum up, if you want to try an offline, local LLM, you can definitely give a shot at Guanaco models.

## 9. Vicuna 33B

Vicuna is another powerful open-source LLM that has been developed by LMSYS. It has been derived from LLaMA like many other open-source models. It has been fine-tuned using supervised instruction and the **training data has been collected from sharegpt.com**, a portal where users share their incredible ChatGPT conversations. It's an auto-regressive large language model and is trained on 33 billion parameters.

In LMSYS's own MT-Bench test, it scored 7.12 whereas the best proprietary model, GPT-4 secured 8.99 points. In the MMLU test as well, it achieved 59.2 points and GPT-4 scored 86.4 points. Despite being a **much smaller model**, the performance of Vicuna is remarkable. You can check out the demo and interact with the chatbot by clicking on the below link.

***Check out Vicuna 33B***

## 10. MPT-30B

MPT-30B is another open-source LLM that competes against LLaMA-derived models. It has been developed by Mosaic ML and fine-tuned on a large corpus of data from different sources. It uses datasets from ShareGPT-Vicuna, Camel-AI, GPTeacher, Guanaco, Baize, and other sources. The best part about this open-source model is that it has a **context length of 8K tokens**.

Additionally, it outperforms the GPT-3 model by OpenAI and scores 6.39 in LMSYS's MT-Bench test. If you are looking for a small LLM to run locally, the MPT-30B model is a great choice.

***Check out MPT-30B***

## 11. 30B-Lazarus

The 30B-Lazarus model has been developed by CalderaAI and it uses LLaMA as its foundational model. The **developer has used LoRA-tuned datasets** from multiple models, including Manticore, SuperCOT-LoRA, SuperHOT, GPT-4 Alpaca-LoRA, and more. As a result, the model performs much better on many LLM benchmarks. It scored 81.7 in HellaSwag and 45.2 in MMLU, just after Falcon and Guanaco. If your use case is mostly text generation and not conversational chat, the 30B Lazarus model may be a good choice.

***Check out 30B-Lazarus***

WizardLM is our next open-source large language model that is built to follow complex instructions. A team of AI researchers has come up with an Evol-instruct approach to rewrite the initial set of instructions into **more complex instructions**. And the generated instruction data is used to fine-tune the LLaMA model.

Due to this approach, the WizardLM model performs much better on benchmarks and users prefer the output from WizardLM more than ChatGPT responses. In the MT-Bench test, WizardLM scored 6.35 points and 52.3 in the MMLU test. Overall, for just 13B parameters, WizardLM does a pretty good job and opens the door for smaller models.

***Check out WizardLM***

## Bonus: GPT4All

GPT4ALL is a project run by Nomic AI. I recommend it not just for its in-house model but to **run local LLMs** on your computer without any dedicated GPU or internet connectivity. It has developed a 13B Snoozy model that works pretty well. I have tested it on my computer multiple times, and it generates responses pretty fast, given that I have an entry-level PC. I have also used PrivateGPT on GPT4All, and it indeed answered from the custom dataset.